AWARD NUMBER: W81XWH-14-1-0234

TITLE: Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers With Lung Cancer

PRINCIPAL INVESTIGATOR: Jennifer Beane-Ebel

CONTRACTING ORGANIZATION: Boston University School of Medicine
Boston, MA 02118-2340

REPORT DATE: July 2016

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
July 2016	Annual	07/01/2015-06/30/2016

4. TITLE AND SUBTITLE

Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers With Lung Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-14-1-0234

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jennifer Beane-Ebel

E-Mail: jbeane@bu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Boston University School of Medicine
Department of Medicine
Division of Computational Biomedicine
Medical Campus
72 East Concord Street, E-631
Boston, MA 02118-2308

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command

Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.

Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. We continue to develop and optimize protocols to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS). We also continue to optimize an adapted version of the CEL-Seq RNA library preparation protocol that we have implemented that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. In addition to computational pipelines to process the sequencing data into gene level counts for each cell, we are developing analysis techniques to define and characterize transcriptionally distinct cell populations. We have successfully sequenced 1,140 cells collected by brushing the bronchial epithelium from 6 never smokers and 6 current smokers. The data reveals the known types of epithelial cells as well as novel subpopulations that have differential abundance in smokers. Genes from our previously developed diagnostic for lung cancer have cell type specific patterns of expression in the single cell data. These results suggest that shifts in transcriptionally distinct cell populations occur in the bronchial epithelium in response to smoking and lung cancer.

Over the next year we plan to sequence between 100 and 200 cell per donor from 12 former smokers and 24 current and former smokers undergoing bronchoscopy for the suspicion of lung cancer. These samples will allow us to characterize the cell populations upon smoking cessation, in high-risk smokers without lung cancer, and in smokers with lung cancer. These data will provide a comprehensive single cell transcriptional profile of changes that occur in the bronchial epithelium in response to smoking, smoking cessation, and lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer.

15. SUBJECT TERMS
single cell, bronchial epithelium, mRNA sequencing, and lung cancer

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	17	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

ABSTRACT:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.

Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. We continue to develop and optimize protocols to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS). We also continue to optimize an adapted version of the CEL-Seq RNA library preparation protocol that we have implemented that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. In addition to computational pipelines to process the sequencing data into gene level counts for each cell, we are developing analysis techniques to define and characterize transcriptionally distinct cell populations. We have successfully sequenced 1,140 cells collected by brushing the bronchial epithelium from 6 never smokers and 6 current smokers. The data reveals the known types of epithelial cells as well as novel sub-populations that have differential abundance in smokers. Genes from our previously developed diagnostic for lung cancer have cell type specific patterns of expression in the single cell data. These results suggest that shifts in transcriptionally distinct cell populations occur in the bronchial epithelium in response to smoking and lung cancer.

Over the next year we plan to sequence between 100 and 200 cell per donor from 12 former smokers and 24 current and former smokers undergoing bronchoscopy for the suspicion of lung cancer. These samples will allow us to characterize the cell populations upon smoking cessation, in high-risk smokers without lung cancer, and in smokers with lung cancer. These data will provide a comprehensive single cell transcriptional profile of changes that occur in the bronchial epithelium in response to smoking, smoking cessation, and lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer.

INTRODUCTION:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract by inducing molecular alterations such as allelic loss, p53 mutations, changes in promoter methylation and telomerase activity[1-5]. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure[6,7]. Importantly, we have extended this airway field of injury to the study of lung cancer, and have identified a bronchial airway gene expression signature that can serve as a diagnostic biomarker for lung cancer[8] that performs independently of clinical risk factors for disease[9]. The lung cancer diagnostic biomarker has been subsequently validated in in a large clinical trial[10] and has been commercialized by Veracyte, Inc. and is known as PERCEPTA™. Advances in technology for amplification of low amounts of RNA combined with next-generation sequencing have produced the ability to characterize the transcriptome of individual cells. While the bronchial brushings examined in our previous studies have captured a relatively pure population of bronchial epithelial cells, we are unable to discern which airway cell type or types are responsible for the gene expression changes observed nor characterize gene expression variation between cells. Variation in gene expression across single cells can be used to define unique subpopulations of cells that may be independent of known markers or cell morphology that may associate with lung cancer. We hypothesize that the lung cancer-specific gene expression in the bronchial epithelium might be restricted to specific known cell types (e.g. basal cells) or molecularly defined subpopulations of cells. The goal of this study will be to use single-cell RNA sequencing to identify cell-type dependent gene expression alterations in the lung cancer field of injury and to molecularly identify novel subpopulations of cells that are associated with lung cancer. These novel molecular insights hold the potential to improve the diagnostic utility of the airway epithelium for lung cancer and to guide new therapeutic strategies for lung cancer prevention.

KEYWORDS: single cell, bronchial epithelium, mRNA sequencing, and lung cancer

ACCOMPLISHMENTS:
What were the major goals of the project?

- Specific Aim 1: Identify which cell types in the airway epithelium harbor the lung cancer-specific alterations in biomarker genes by single cell RNA sequencing
    - Major Task 1: Isolate and sequence the RNA of single epithelial cells from the bronchus of smokers with and without lung cancer (n=15 subjects/group, n=960 single cells/subject).
        - Subtask 1: Approval of IRB and HRPO (1-2)

The IRB at Boston University School of Medicine approved the study protocol on September 10, 2014 but notification of the outcome was received on December 4, 2014. The HRPO approval

was obtained on December 19, 2014. This process took approximately 6 months to complete and delayed sample collection by about 4 months. Completion percent: 100%

- Subtask 2: Collection of airway brushings from 30 subjects at BUMC (24-30)

  *We have collected airway brushings from 28 never, current, and former smokers and 15 current and former smokers with and without lung cancer undergoing bronchoscopy for clinical suspicion of lung cancer. Of those subjects, 6 have been confirmed to have lung cancer and 9 do not have lung cancer (**Table 1**). We will continue to collect over the next year in hopes of collecting samples from more subjects with cancer. Completion Percent: 85%*

- Subtask 3: Sorting of cells from brushings using FACS (Sorting of cells will take place within hours after collection) (24-30)

  Tissue acquired via bronchial brushings is dissociated, dead cells and red blood cells are excluded via FACS, and live cells are sorted into 96-well plates and frozen. This process needs to be completed immediately after sample collection. Therefore, we have sorted all samples collected above and completion is dependent on sample collection.
  Completion Percent: 85%

- Subtask 4: mRNA isolation and library preparation (24-30)

  An established technique (CEL-Seq) for preparing single cell RNA-Seq libraries was adapted for this project and modified to increase sample multiplexing capacities and correct for experimental amplification biases. To date, we have produced high quality single cell data using this protocol and in the coming year we will process the samples collected. Completion Percent: 85%

- Subtask 5: Sequencing of samples on Ilumina HiSeq 2500 (26-32)

  Single cell RNA libraries are massively multiplexed and paired-end sequencing is performed using the Illumina HiSeq 2500. Currently we have sequenced cells from 12 of the total 43 samples collected. Completion Percent: 28%

- Subtask 6: De-multiplex samples, preprocess, align, and analyze data quality (32)

  A computational pipeline developed in collaboration with the Yanai Lab (http://yanailab.technion.ac.il/) was used to preprocess and align reads generated via the CEL-Seq methodology. Additional metrics have been incorporated for the purposes of quality control and determination of sample cell type. The pipeline developed will be used to process the data as it is generated. In addition, we spike-in ERCC controls into each well and compare the number of reads aligned to these controls to their known concentrations. A high correlation between read count and ERCC control concentration is on indicator of quality. Percent complete: 28%

- Milestone(s) Achieved: Generation of high quality single cell RNA sequencing data from 30 subjects

  After years of work, we have successfully created an experimental protocol that yields high-quality single cell sequencing data. Using this protocol, we have successfully generated high-quality data on 1,140 cells from 12 subjects. The data provides unprecedented insight into the cell populations present in the bronchial airway epithelium. We will use this protocol over the next year to complete sequencing the cells collected and sorted for an additional 31 subjects.

o Major Task 2: Determine the cell type(s) responsible for the aberrant gene expression in the airway lung cancer diagnostic biomarker. Completion Percent: 75%

- Subtask 1: Summarize sequencing data into counts per gene (30)

  We have implemented a computational pipeline to summarize the sequencing data into counts per gene. We have gene counts for all 1,140 cells sequenced to date. Percent complete: 28%

- Subtask 2: Classify gene expression as signal or noise based on a mixture model (30)

  An average of ~1500 genes were detected per cell and genes with less than 2 detected transcripts in 5 cells were excluded from further analyses. Percent complete: 28%

- Subtask 3: Determine cell types of origin for lung cancer biomarker genes (30-34)

  We have defined transcriptionally distinct cell populations using Latent Dirichlet Allocation (LDA), a form of mixture modeling that can be used to determine the probability that sets of co-expressed genes (interpreted as "transcriptional states") are expressed by specific cellular subsets. Populations and their corresponding transcriptional states were visualized using the t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction approach. We have conducted this analysis using the data from the 1,140 cells. We have subsequently examined expression of our lung cancer biomarker genes in these populations. Percent complete: 28%

- Subtask 4: Identify cell type dependent lung cancer associated differential expression using a linear modeling and ANOVA strategy (32-34)

  Cellular subclassification was performed using t-SNE and density clustering, which led to identification of 6 groups of cells (basal, secretory, ciliated, neuroendocrine, and 2 sets of white blood cells) [data not shown]. A rank-based (KS) statistical test was used to determine enrichment of LDA-derived "transcriptional states" amongst identified cell clusters as well as amongst current smoker cells ("Cell Set Enrichment Analysis"). Furthermore, genes with high probabilities of being affiliated with each LDA-derived "transcriptional state" were assigned to each enriched cell type (an example of the secretory cell-specific "transcriptional states" and their corresponding genes is displayed in Figure 3). We will implement these same strategies to identify lung cancer-associated cell type-specific genes over the next year as we process the collected samples. Percent complete: 28%

- Subtask 5: Choose candidates for validation (34)

  We will be obtaining normal biopsies from the bronchial airway from current and former with and without cancer. We plan validate our findings using immunofluorescence to stain for expressed by lung cancer specific cell populations.

- Milestone(s) Achieved:
  We have identified known lung cancer specific gene expression alterations that we observe to have cell type specific expression upon projection into our single cell data. Sequencing of samples from lung cancer subjects over the next year will confirm these observations and likely identify additional genes and cell populations.

- Specific Aim 2: Identify unique cell populations in the airway of smokers that are associated with lung cancer. Completion Percent: 45%
  - Major Task 1: Molecularly identify subpopulations of cells irrespective of cell type and determine if these cells are more or less abundant in the airways of patients with lung cancer
    - Subtask 1: Identification of novel subpopulations of cells using both class discovery and pathway prediction approaches (34-36)

      We have used LDA to identify transcriptionally distinct cell populations and genes expressed by these populations and are characterizing the functions (pathways) of each population. We will implement these same strategies to identify lung cancer-specific populations of cells over the next year as we process the collected samples.

    - Subtask 2: Identify subpopulations associated with lung cancer

      We have identified smoking-associated subpopulations using LDA. We will implement these same strategies to identify lung cancer-specific subpopulations over the next year as we process the collected samples.

    - Subtask 3: Choose candidates for validation (34-36)

We are choosing various cell population specific markers to validate via immunofluorescence in using cytospins of bronchial epithelial cells collected during bronchoscopy from additional subjects that we are currently recruiting. This work is a foundation for how we will validate the finding in the subjects with cancer.

- Milestone(s) Achieved: Identification of novel subpopulations of airway epithelial cells that are associated with lung cancer (34-36)

  We have identified subpopulations of airway epithelial cells in 1,140 cells collected from never and current smokers. We have characterized the expression patterns of our previously defined lung cancer-specific biomarker genes in this dataset. The analysis reveals genes specific to secretory cells are consistently down in cancer, whereas genes specific to ciliated cells are consistently up. As we run more samples we will be able to enhance our understanding of how these populations change present upon smoking cessation and in the presence of lung cancer.

o Major Task 2: Validate lung cancer associated genes from specific cell types or from novel subpopulations in bronchial epithelial cells from independent subjects (n=10) using FISH
  - Subtask 1: Collect airway brushes from 10 subject for validation (24-36)
  - Subtask 2: RNA-FISH will be used to validate 5 candidate genes in conjunction with 4 known epithelial marker genes (30-36)
  - Milestone(s) Achieved: Validation of novel lung cancer-associated gene candidates in specific populations of epithelial cells

What was accomplished under these goals?

The major objective is to conduct single cell RNA sequencing on cells collected from bronchial brushings from current and former smokers with and without lung cancer. In order to attain this goal, we developed and optimized the experimental protocols for FACS sorting, library preparation, and single cell data analysis. We have developed an unbiased methodology for sorting the cells from bronchial brushings into 96-well plates, a library preparation protocol that will produce high quality data, and an analysis pipeline for processing the data. As a result of the significant challenges in producing high-quality data, we elected to first study bronchial epithelial cells from never and current smokers. This follows our previous work of characterizing smoking-associated gene expression changes in the bronchial airway epithelium prior to our focus on lung cancer[6,7]. The smoking-associated gene expression changes found by microarray in our previous studies are very robust and reproducible and represented an ideal system for optimizing our protocol. We have generated high-quality data on 6 never and 6 current smokers by sequencing 1,140 cells. This experiment has revealed transcriptionally distinct populations of epithelial and immune cells. Smoking is associated with distinct shifts in these cell populations as well as differential expression within these populations between never and current smokers. In order to accomplish the goals outlined in the original proposal we have been collecting and sorting brushes from former smokers and current and former smokers with and without lung cancer. We plan to run all of these banked samples within the no cost extension (NCE) period of this grant to produce results in the context of lung cancer. The current results, however, although in never and current smokers provide important insights into lung cancer. A subset of our previously defined lung cancer-specific biomarker genes are expressed in specific cell populations and may indicate important cell type shifts that occur with the onset of lung cancer that could be potentially targeted in high-risk smokers.

Sample Collection

Current and former smokers were recruited that were undergoing flexible bronchoscopy for clinical suspicion of lung cancer at BUMC. For each consented subject, we collect data regarding their age, gender, race, and a detailed smoking history. Additional samples were collected from healthy never, current and former smoker volunteers recruited for a project entitled "Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products" (**Table 1**). The subjects in this study were recruited via advertisements, a two-part screening questionnaire was administered via telephone to determine eligibility, and eligible subjects completed two study visits, the second of which included a bronchoscopy. These healthy donors will be treated as controls and data generated for the "Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products" project will be used in conjunction with data

Cohort	Smoking status	Cancer status	Age (SD)	Sex	Race
Clinical suspicion of lung cancer	7 current smokers	4 cancer	65 (2)	4 F	4 A.A.
		3 no cancer	60 (10)	3 F	2 C. / 1 A.A.
	8 former smokers	2 cancer	78 (11)	1 M / 1 F	2 C.
		6 no cancer	70 (10)	4 M / 2 F	4 C. / 2 A.A.
Healthy volunteers	11 current smokers	NA	41 (12)	7 M / 4 F	5 C. / 6 A.A.
	10 former smokers	NA	41 (11)	3 M / 7 F	7 C. / 2 A.A. / 1 As.
	7 never smokers	NA	29 (7)	4 M / 3 F	3 C. / 2 A.A. / 2 As.

Table 1. Samples collected to date.

C.* = Caucasian
A.A.* = African American
As.* = Asian

generated from samples collected for this project.

Collection of bronchial brushings is done using the same technique as in our prior studies[6,8]. Following topical anesthesia of the upper airway using 2% lidocaine, a bronchoscope is introduced to the right mainstem bronchus and epithelial cells are obtained using an endoscopic cytobrush.

Cell sorting by FACS

Tissue obtained from bronchial brushings is treated with 0.25% Trypsin/EDTA for epithelial sheet dissociation and cells are sorted using a BD FACSAria II. Gating based on forward scatter height vs. forward scatter area (FSC-H vs. FSC-A) is applied to sort only singlet events (single cells). Staining for GYPA (CD235a) is used to exclude all red bloods cells. Hoechst 33342 is used to stain the DNA of all cells, whereas PI is used to specifically stain the DNA of dead cells with compromised membranes. For each donor, single Hoechst 33342^{+} PI^{-} CD235a^{-} cells are sorted into five 96-well PCR plates (480 cells), frozen on dry ice and stored at -80 °C until preparation for sequencing (**Figure 1**).

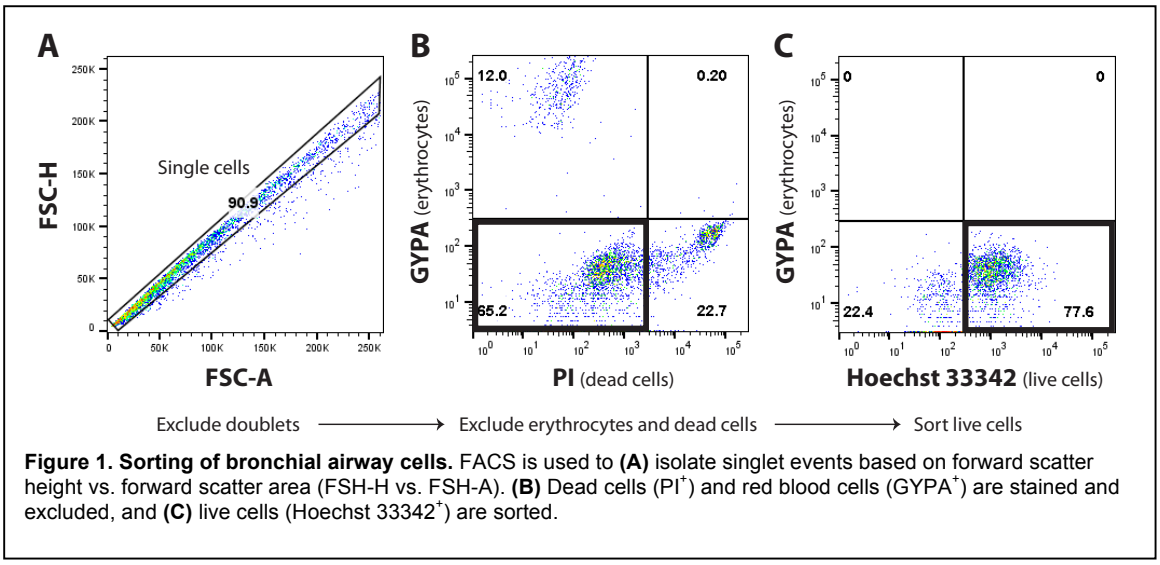

Figure 1. Sorting of bronchial airway cells. FACS is used to **(A)** isolate singlet events based on forward scatter height vs. forward scatter area (FSH-H vs. FSH-A). **(B)** Dead cells (PI^{+}) and red blood cells (GYPA^{+}) are stained and excluded, and **(C)** live cells (Hoechst 33342^{+}) are sorted.

Development of a single cell mRNA sequencing protocol

Massively parallel single cell RNA-sequencing of human bronchial airway cells is being performed using the CEL-Seq RNA library preparation protocol that has been modified for increased high throughput capacities[11]. Frozen 96-well PCR plates containing sorted cells are thawed on ice and RNA is directly reverse transcribed (Ambion AM1751) from whole cell lysate using primers composed of an anchored poly(dT), the 5' Illumina adaptor sequence, a well-specific barcode, a random sequence, which serves as a transcript-specific unique molecular identifier (UMI)[12], and a T7 RNA polymerase promoter. Samples are additionally supplemented with ERCC RNA Spike-In mix (Ambion) for quality control[13]. cDNA generated from each of the 96 cells per plate is uniquely barcoded and therefore can be pooled for second strand synthesis (Ambion AM1751) and amplification by in vitro transcription (Ambion AM1751). Amplified RNA is then chemically fragmented (NEB E6150) and ligated to the Illumina RNA 3' adapter (Illumina RS-200-0012). Samples are again reverse transcribed and amplified using indexed Illumina RNA PCR primers (Illumina RS-200-0012). This barcoding strategy allows for the loading of libraries derived up to 4608 cells (96 well-specific barcodes x 48 plate-specific indices) onto a single flow cell lane for paired-end sequencing using the Illumina HiSeq 2500.

Development of a computational pipeline to process sequencing data

In collaboration with the Yanai lab (http://yanailab.technion.ac.il/), we utilized a computational pipeline to preprocess and align reads generated from the CEL-Seq protocol. Briefly, reads were demultiplexed into "plate" level FASTQ files using Illumina software. Each plate FASTQ file was further demultiplexed into cell specific FASTQ files using custom scripts. The UMI barcode was also removed from each read and stored in the read header. Reads were aligned to the human genome using Bowtie2. Gene level counts were derived using a modified version of the HTSeq python library, which only counts multiple reads that are aligned to the same position and have the same UMI once, thus eliminating PCR amplification bias. Quality of each cell was assessed by examining the total number of reads aligned, the total number of genes detected, the number of ERCC spike in transcripts detected, the correlation between read counts and the ERCC transcript concentrations, and known airway cell type markers.

Pilot Sequencing Experiment

In the last report, we presented a pilot experiment using bronchial epithelial cells from 1 healthy former smoker and 1 healthy current smoker (n=24 cells per donor) were profiled (48 total cells) via paired end sequencing using the Illumina MiSeq. Epithelial cells were sorted using the FACSAria II, according to expression of the established airway epithelial surface marker CD166[14]. Expression of major epithelial subpopulation markers was detected in subsets of cells profiled (KRT5, MUC5AC, SCGB1A1, FOXJ1); however, upon further data analysis we determined that SCGB1A1, a marker of Clara secretory cells, was expressed at high levels across all cell types. These results indicated a technical library preparation-related issue. After troubleshooting the protocol, we determined that the 96-well format nucleic acid purification protocol was leading to inadvertent well-to-well cross contamination. This protocol was altered such that the nucleic acid purification step was excluded entirely and replaced with a direct-to-RT (reverse transcription) approach. This alteration largely resolved the cross contamination problem, resulting in the production of a higher quality dataset and increased capacity to adequately interpret cell type-specific gene expression.

Single Cell RNA-Seq of the Bronchial Epithelium of Never and Current Smokers

After resolving the contamination issue, we conducted a larger experiment to assay 96 cells from 12 subjects. Specifically, bronchial brushings were obtained from 6 never smokers and 6 current smokers. For each donor, single ALCAM^{+} epithelial cells or CD45^{+} white blood cells (WBCs) were sorted into 96-well PCR plates containing while excluding red blood cells and dead cells. Single cell RNA-seq of human bronchial airway cells was performed using the CEL-Seq RNA library preparation protocol and sequenced on the Illumina HiSeq 2500[9]. ERCC spike-in RNA was put into every well to serve as a positive control and one well was left empty as a negative control. After aligning all reads to the genome for each cell we quantified the expression of all protein-coding genes. We used a clustering approach called t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the relationship between cells (**Figure 2**). Never and current smoker cells were highlighted in grey and black, respectively (**Figure 2A**). Furthermore, relative expression of known airway epithelial marker genes was denoted in red (KRT5 = basal cells) (**Figure 2B**), yellow (FOXJ1 = ciliated cells) (**Figure 2C**), green (MUC5AC = goblet cells) (**Figure 2D**), and blue (SCGB1A1 = Clara cells) (**Figure 2E**). Visual inspection of these plots indicates that there are subsets of basal, ciliated, goblet, and Clara cells are clearly present. More so, goblet cells appear to be more abundant in current smokers, whereas basal and Clara cells are more abundant in never smokers.

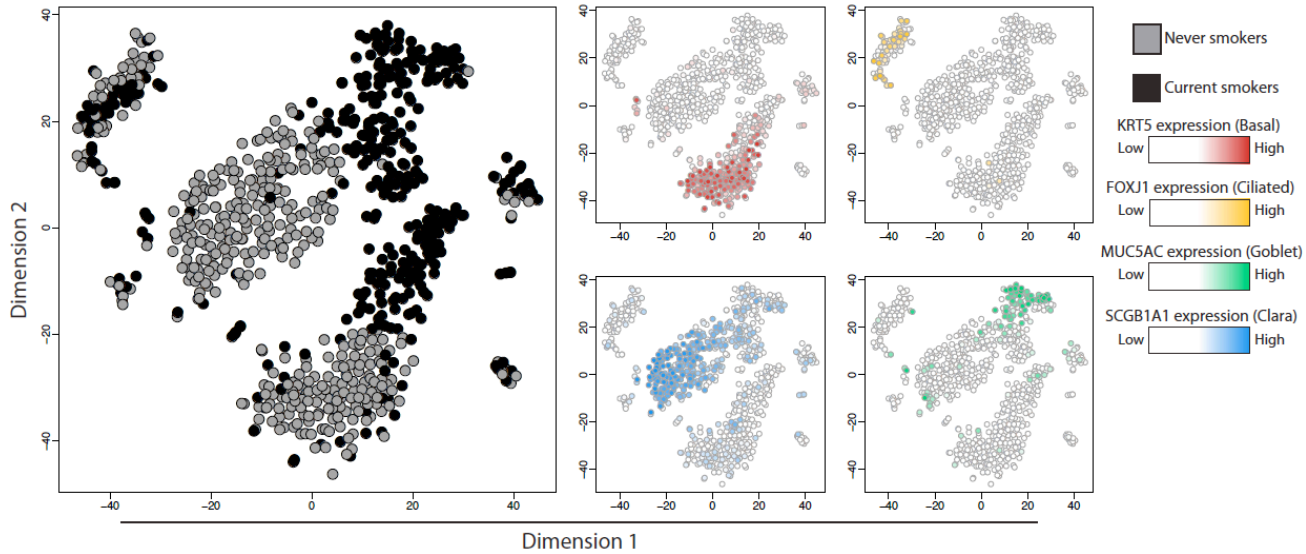

Figure 2. Single cell RNA-seq of normal bronchial epithelium. We performed single cell RNA-seq on 1,140 cells collected from 6 current smokers and 6 never smokers. Cells were clustered across all genes using t-SNE. Cells with similar global gene expression patterns will be in similar locations on the plot. Each point represents a single cell and each plot shows the same clustering of cells. **(A).** Cells are colored by smoking status where black cells are from smokers and grey cells are from nonsmokers. **(B-E)** Cells are colored by the expression levels of known marker genes for different cell types including basal, ciliary, and mucin-producing (goblet/Clara). Overall, these plots demonstrate that cells tend to cluster by both cell-type and by smoking status and that smoking can induce cell-type specific changes in response to smoke, a finding previously unappreciated by bulk profiling of this tissue.

Additionally, we have defined transcriptionally distinct cell populations using Latent Dirichlet Allocation (LDA), a form of mixture modeling that learns transcriptional states from a collection of cells and identifies which genes are likely to belong to that state. In order to visualize the states we cluster the cells across all genes using t-SNE and color the cells according to their probability of membership in each transcriptional state. Our analysis revealed 14 distinct states that divide the known cell types illustrated in **Figure 2** into transcriptionally distinct subpopulations. **Figure 3** illustrates the 6 transcriptional states that characterize the secretory cells profiled. States 14 and 5 are more specific to cells from never

smokers while states 12, 4, and 10 are more specific to cells from current smokers. Genes with high probabilities of belonging to each state are illustrated in heatmaps. Our previous work profiling bronchial epithelial cells from current, former, and never smokers using gene expression microarrays demonstrated that upon smoking cessation gene expression levels return to levels similar to never smokers with different kinetics[7]. Genes classified as rapidly reversible return to levels similar to never smokers in months; however, slowly reversible genes take several years. Enrichment of slowly reversible genes in transcriptional state 14, for example, suggests that this subpopulation of secretory cells present in the airway of never smokers may be under-represented upon smoking cessation. Similarly for transcriptional state 10, the subpopulation of goblet secretory cells that are present in the airway of smokers may persist upon smoking cessation. Additional single cell profiling of collected samples from former smokers will confirm these observations. We can hypothesize that these two populations of cells may indicate risk of developing disease.

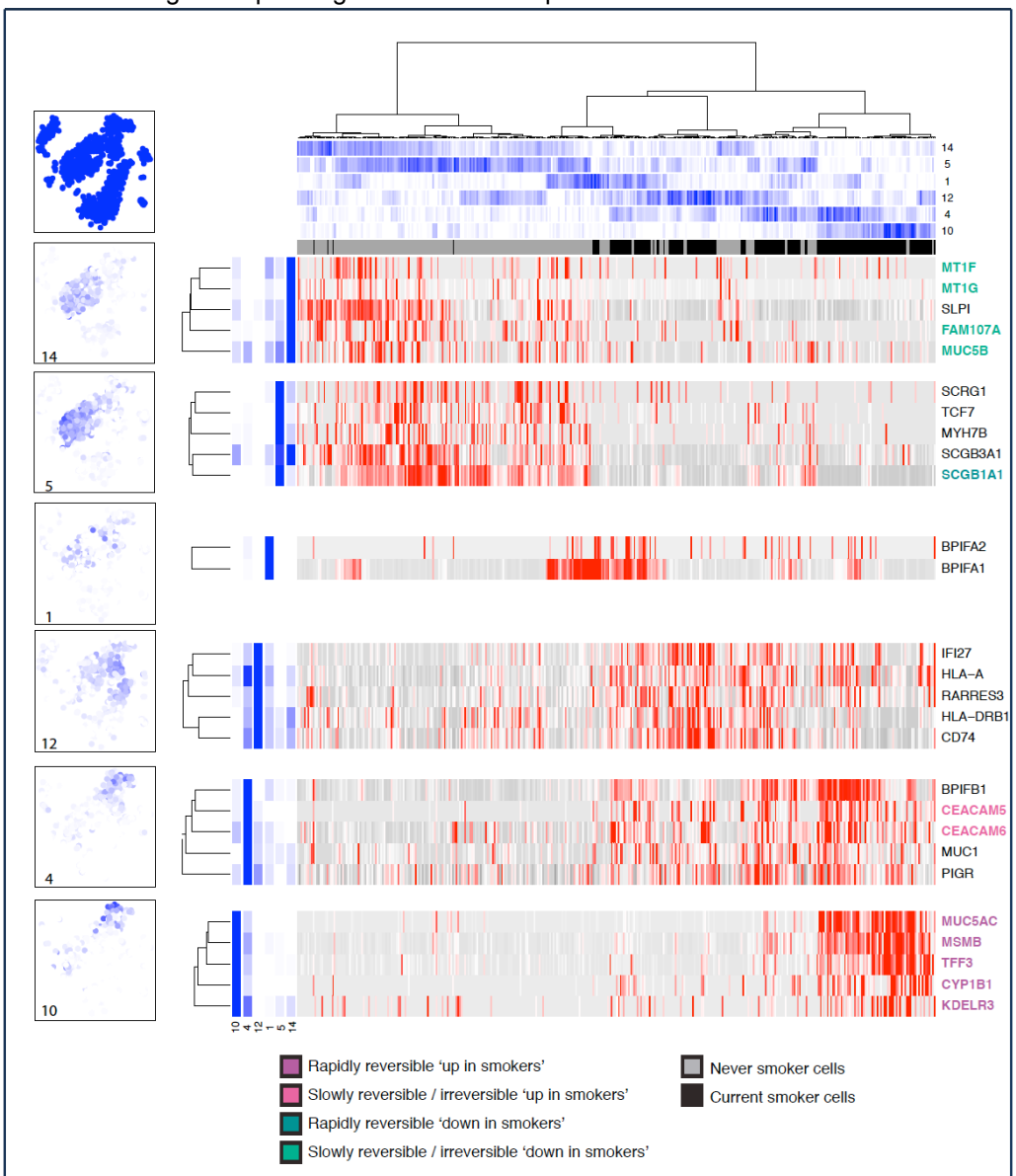

Figure 3. LDA Analysis illustrating secretory cell transcriptional states. LDA analysis was performed across the single cell RNA-seq data of 1,140 cells collected from 6 current smokers and 6 never smokers. Left panel illustrates cells clustered across all genes using t-SNE where each transcriptional state (14, 5, 1, 12, 4, or 15) is colored by the probability that the cell belongs to that state. Each point represents a single cell and each plot shows the same clustering of cells. Blue indicates high probability and white indicates low probability. The t-SNE plot for each state corresponds to a heatmap in the right panel that clusters all cells according to representative genes with high probabilities of belonging to each of the transcriptional states. States 14 and 5 are more specific to cells from never smokers while states 12, 4, and 10 are more specific to cells from current smokers. Gene names are colored by the rate at which expression returns to levels similar to never smokers upon smoking cessation from our previous work.

Expression of Lung Cancer Biomarker Genes in the Single Cell RNA-Seq Data

We have previously described lung cancer-specific gene expression changes in the airway field of injury that have been leveraged to build a gene expression-based diagnostic for lung cancer[8,10]. The diagnostic has been commercialized by Veracyte and is marketed as a test called PERCEPTA™. The goal of this grant to understand if any of these gene expression changes were cell type specific and to identify novel subpopulations of cells within and across cell types that harbor lung cancer-associated gene expression alterations. In the no cost extension period, we will profile cells from current and former smokers with and without lung cancer to try and answer these questions; however, we can leverage the single cell data in hand to identify genes that are expressed in a cell type-specific manner. We examined the expression of the 232 genes reported to be associated with lung cancer in the bronchial airway[15] in the single cell data shown in **Figures 2 and 3**. We found that a subset of these genes have cell type-specific gene expression (**Figure 4**). Interestingly, genes (LYPD2 and PI3) are specifically expressed by secretory cells found in never smokers and are down-regulated in lung cancer our previous work[15]. Similarly, LCN2, that is expressed by never and current smoker secretory cells is also down-regulated in lung cancer. Expression of NKX3-1 is more specific to the goblet-like secretory cells of current smoker and is up-

regulated in lung cancer. NKX3-1 is a tumor suppressor and is up-regulated in these cytologically normal bronchial epithelial cells collected from the right mainstem bronchus suggesting that is may be important in maintaining normal epithelium[16]. Three of the genes, SLC7A2, VSTM2L, and CES1, are specific to ciliated cells and are up-regulated in lung cancer. In the current single cell dataset, we can identify gene expression differences between current and never smokers among ciliated cells, suggesting that there will likely be similar differences between subjects with and without lung cancer. The cilia in high-risk smoking subjects may be impaired and signal an up-regulation of ciliogenesis.

What opportunities for training and professional development has the project provided?
Grant Duclos, a graduate student, responsible for sorting the cells and preparing the libraries as part of the project has had the opportunity to study under Dr. Itai Yanai, Associate Professor at Technion – Israel Institute of Technology. Dr. Yanai is on sabbatical at the Broad Institute and has extensive experience in single cell sequencing and under his mentorship Grant has been able to effectively troubleshoot our protocols. Additionally, during the period of this grant, Joshua Campbell has become an Assistant Professor at BUMC in the Department of Medicine and Section of Computational Biomedicine. Using the knowledge he has gained over the course of this project he started a Single Cell

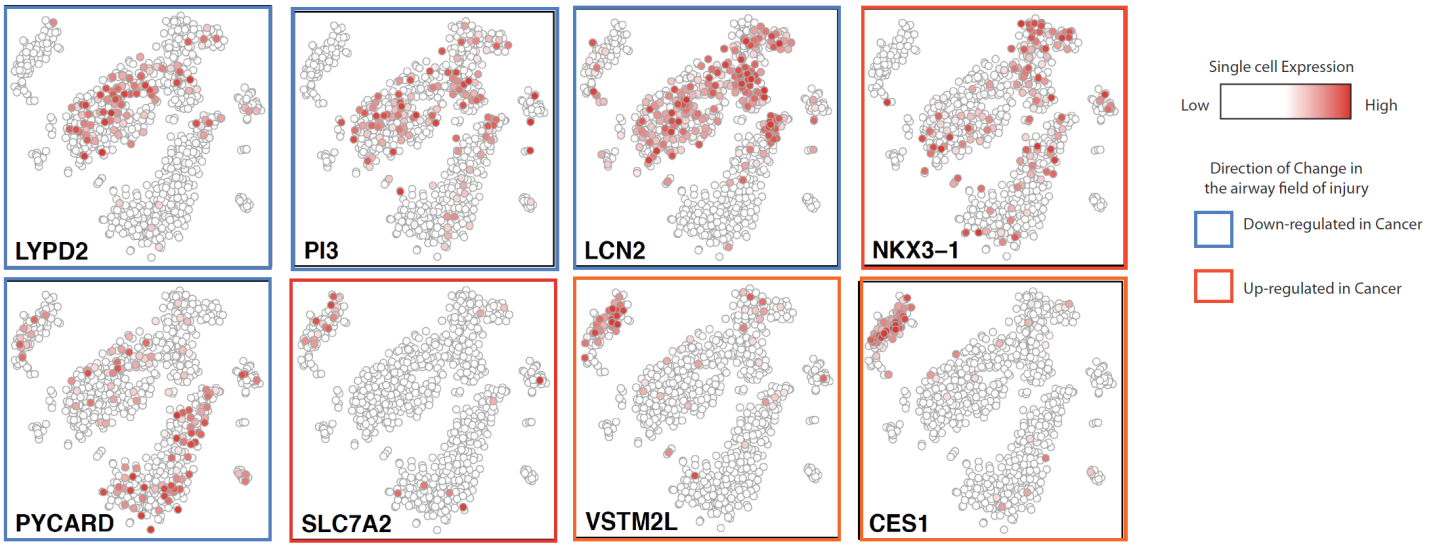

Figure 4. Expression of Select Lung Cancer-Specific Genes in Single Cells from Current and Never Smokers. Single cell RNA-seq on 1,140 cells collected from 6 current smokers and 6 never smokers was clustered across all genes using t-SNE. Each point represents a single cell and each plot shows the same clustering of cells. The colors indicate expression of the gene in the single cells, with a red color indicating higher expression. The colored boxes (red or blue) around each t-SNE plot indicate the direction of change in bronchial epithelial cells from lung cancer subjects profiled in bulk on microarrays compared to subjects without cancer.

Working Group at BUMC that attracts researchers with diverse interests. The Single Cell Working Group is designed to educate the community about single cell sequencing and to discuss library preparations protocols and data analysis techniques. The goal is to develop a single cell sequencing analysis toolkit. In addition, these study are leading to the development of a single cell tumor bank by the Cancer Center at BUMC.

How were the results disseminated to communities of interest?
The results of this work were presented at the American Thoracic Society Annual Meeting between May 13th and 18th, 2016 in San Francisco, CA in a Poster Discussion Session. In addition, Jennifer Beane gave a talk at the National Cancer Institutes Annual Lung SPORE meeting between June 23rd and 24th, 2016. The references and abstracts for these presentations is below:

- Duclos GE, Campbell JD, Autissier P, Dumas YM, Terrano R, Liu G, Lenburg ME, Spira A, Beane J. Single Cell RNA Sequencing Reveals Smoking-Associated Alterations in Bronchial Epithelial Subpopulations. Poster A4629. American Thoracic Society Annual Meeting. 2016.

  RATIONALE: Bronchial epithelial gene expression reflects the physiologic response to cigarette smoke exposure. Our group has described transcriptomic alterations in the airways of current smokers as well as alterations that persist after smoking cessation. In this study, we use single cell mRNA sequencing to profile the transcriptomes of individual bronchial epithelial cells obtained via bronchial brushing from current and never smokers. This could allow for the detection of smoking-associated alterations within specific epithelial cell types, shifts in cell type

abundance in smokers, as well as the potential for discovery of novel cell types affected by tobacco exposure.
METHODS: We obtained bronchial brushings from current smokers (n=6) and never smokers (n=6) and isolated single epithelial cells by FACS. The CEL-Seq RNA library preparation protocol was used to simultaneously sequence the transcriptomes of 1,008 cells (n=84/donor). Single cell library quality was assessed using ERCC controls. Cellular subpopulations were identified according to the expression of known proximal airway epithelial marker genes (KRT5, MUC5AC, and FOXJ1). Negative binomial generalized linear models were used to identify genes whose expression was significantly associated with smoking status (FDR $q < 0.05$) within specific cell types. Sets of differentially expressed (DE) genes were functionally annotated using the GO Biological Process database and the web-based tool suite Enrichr and intersected with our previously published smoking-associated bronchial gene expression signature.
RESULTS: Libraries of high quality were analyzed (n=990). Decreased numbers of KRT5+ and FOXJ1+ cells were observed in current smokers, while numbers of MUC5AC+ cells were increased. The gene, MUC5AC, was also up-regulated in MUC5AC+ cells from smokers. Many of the smoking-associated gene expression differences previously identified in unfractionated bronchial brushes were found in different cellular subsets. Functional annotation results indicate that genes up-regulated in smoker MUC5AC+ and KRT5+ cells are involved with antigen presentation and cytoskeletal organization, respectively, and genes down-regulated in smoker FOXJ1+ cells are involved with protein folding.
CONCLUSION: These results suggest that the gene expression response to smoking detected in bronchial airway brushings is an ensemble of different responses from multiple cell types within these samples. Findings from further single cell RNA-seq studies in this setting will allow us to better interpret the molecular consequences of smoking on the bronchial epithelium and may provide insight into how smoking contributes to disease pathogenesis.

- Duclos GE, Campbell JD, Autissier P, Dumas YM, Terrano R, Liu G, Lenburg ME, Spira A, Beane J. Single Cell RNA Sequencing Reveals Smoking-Associated Alterations in Bronchial Epithelial Subpopulations. Talk at National Cancer Institute's Annual Lung SPORE meeting. 2016.

  Objectives: Bronchial epithelial gene expression reflects the physiologic response to cigarette smoke exposure. Our group has described alterations in the airway transcriptome associated with smoke exposure as well as the dynamic changes that occur with smoking cessation by profiling airway brushes obtained during bronchoscopy. In this study, we profile bronchial epithelial cells obtained from current and never smokers using single cell mRNA sequencing to characterize the response to smoke exposure. In contrast to profiling of the bulk cell population, the single cell resolution is able to detect smoking associated alterations in epithelial subpopulations as well as shifts in subpopulation abundance.
  Methods: We obtained bronchial brushings from current smokers (n=6) and never smokers (n=6) and isolated single epithelial cells by FACS. The CEL-Seq low input RNA library preparation protocol was used to simultaneously sequence the transcriptomes of 1,152 cells (n=96 cells/donor). Dimensionality reduction was conducted using t-distributed stochastic neighbor embedding and cellular subpopulations were defined using density-based clustering. Negative binomial generalized linear models were used to identify genes differentially expressed in each subpopulation (FDR $q < 0.05$). Sets of differentially expressed (DE) genes were intersected with our previously published smoking-associated bronchial gene expression changes.
  Results: Distinct subpopulations of bronchial cells were identified and cluster-specific signatures were generated to determine predominant cell types in never smoker samples (primarily basal, secretory, and ciliated epithelial cells). Cells from smokers were transcriptionally distinct compared to never smokers between related subpopulations and within subpopulations. Smoking induces epithelial subpopulation-specific gene expression signatures related to dysregulated oxidative phosphorylation, IFNg signaling, the unfolded protein response, and xenobiotic metabolism. There is an expanded presence of $KRT8^{+}$ epithelial progenitors and $MUC5AC^{HIGH}$ and $SCGB1A1^{Low}$ secretory cells in the airways of smokers. The findings also indicate that the previously published 'whole tissue' smoking signature derived from airway brushings is comprised of multiple cell type-specific smoking signatures.
  Conclusions: Single cell RNA-seq identified airway epithelial subpopulation-specific transcriptomic alterations and shifts in epithelial subset abundance associated with smoking. Findings from this study will allow us to better interpret the molecular consequences of smoking on the bronchial epithelium and may provide insight into how smoking contributes to disease pathogenesis.

What do you plan to do during the next reporting period to accomplish the goals?
We have recently hired a new technician to process the banked samples. In the next few months after his training is complete he will be able to process the remaining of samples that we have collected and continue to collect. Once these samples are processed and data is available, we will be able to assess single cell gene expression changes in response to smoking cessation and the presence of lung cancer. In addition, we are in the process of validating the data

presented in this report and are preparing a manuscript that will be published in the next year.

IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

To date, human bronchial epithelial cells obtained via bronchoscopy have not been sorted and processed for single cell RNA sequencing. The cells need to be immediately taken from the bronchoscopy suite and processed for FACS sorting. FACS sorting of the cells is accomplished as stated above and the sorted cells are frozen in 96-well plates. The CEL-Seq RNA preparation protocol has been modified to provide the ability to process hundreds of cells from several subjects. All of the methodology and protocols developed as part of this project will directly benefit other groups that are attempting to examine single cell transcriptomics in human clinical samples. In addition, all microarray or RNA sequencing studies performed to date on human airway epithelial cells have been conducted using a bulk population of cells. Expression values represent the mean behavior of all the cells profiled in a given sample and do not capture cell-to-cell gene expression variation. Using a single cell approach, we are able to characterize the cell types (both known and novel) that are present and begin to identify the cell types responsible for the lung cancer-specific airway field of injury. From the data presented above, the diversity of secretory cell populations (**Figure 3**), the presence of previously identified genes that do not return to never smoker gene expression levels for several years (slowly reversible genes, **Figure 3**) in a subset of these secretory cell populations, and the expression of lung cancer specific genes in secretory cells (**Figure 4**) suggest that these cells will be important in the lung cancer field on injury. Using our sorting technique, we will also be profiling immune cells present when the brushing was obtained. In summary, the study contributes to our understanding of the human airway epithelium and the changes that occur in response to smoke exposure and the acquisition of lung cancer.

What was the impact on other disciplines?

The project is multidisciplinary and will impact the study of lung cancer and epithelial cell biology as well as contribute to molecular biology and bioinformatics methods. The current data presented above suggests new markers to study epithelial cell types/subpopulations that appear to be important in the process of lung carcinogenesis. Additionally, the application of computational techniques such as LDA that are being developed as part of this project will be useful to researchers analyzing other single cell datasets.

.

What was the impact on technology transfer?

Nothing to report.

What was the impact on society beyond science and technology?

A gene expression based biomarker for improving lung cancer diagnosis known as PERCEPTA^{TM} and commercialized by Vercyte (http://www.veracyte.com) is currently available following promising clinical trial results[10]. The test helps identify patients at low risk for having lung cancer after a non-diagnostic bronchoscopy ordered as a result of CT scan abnormalities. The findings in this project may help develop an improved diagnostic test that will impact the clinical management of high lung risk patients.

CHANGES/PROBLEMS:

Changes in approach and reasons for change

As stated in our previous progress report, we expanded sample collection to include bronchial brushings collected from healthy current and former smoker volunteers recruited for a project entitled “Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products”. These healthy donors were critical to establishing the CEL-Seq library preparations protocols, as we have had to troubleshoot the protocol extensively to generate high-quality data. The data from these donors represents a valuable set of control samples that we can use to increase our understanding how the bronchial airway epithelium changes in response to smoking. Smoking-associated gene expression changes in our previously published work are very robust in comparison to changes associated with lung cancer. In the data presented, a substantial shift in the cell types present is observed and many genes overlap with those previously published. We will be able to compare the single cell results obtained from smokers with lung cancer to these data to elucidate shifts in cell type or cell type specific expression changes that occur in subjects with lung cancer.

Actual or anticipated problems or delays and actions or plans to resolve them

To date, Grant Duclos, a graduate student has been responsible to collecting, sorting, and processing the samples. Grant has also been analyzing the data under the guidance of Joshua Campbell and Jennifer Beane. Grant doesn’t have the capacity to process the banked samples although he will continue to collect and sort samples for single cell sequencing and validation studies. A new technician is being trained to process the samples and we believe this will overcome the delay in data generation. Finally, collections of samples at Boston University Medical Center has been slow; however, we believe we have accrued enough samples to publish papers on smoking-associated alterations, smoking cessation-associated alterations, and lung cancer-specific alterations. Additionally, as part of these studies we hope to leverage the thousands of bulk bronchial epithelial cell samples that have been profiled by microarray and standard RNA-sequencing

in the laboratory of Dr. Avrum Spira by using the single cell type signatures to deconvolute the bulk profiled tissue.

Changes that had a significant impact on expenditures
Nothing to Report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals.
Nothing to report.

Significant changes in use of biohazards and/or select agents
Nothing to report.

PRODUCTS:
Publications, conference papers, and presentations
Nothing to report.

Journal publications.
Nothing to report.

Books or other non-periodical, one-time publications.
Nothing to report.

Other publications, conference papers, and presentations.
Nothing to report.

Website(s) or other Internet site(s)
Nothing to report.

Technologies or techniques
The techniques, including cell sorting, library preparation, and analysis methods developed in this project will be shared through publication of a manuscript reporting the findings of single cell sequencing experiments profiling smokers with the without lung cancer.

Inventions, patent applications, and/or licenses
Nothing to report.

Other Products
Nothing to report.

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS
What individuals have worked on the project?

Name:	*Jennifer Beane-Ebel*
Project Role:	*Principal Investigator*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*3.0*
Contribution to Project:	Overseeing all aspects of the project including sample collection, experimental design, and data analysis
Funding Support:	DOD IDA, LUNGevity Foundation, NIH/NIAID, Industry award, Internal funds

Name:	*Joshua Campbell*
Project Role:	Postdoctoral Fellow
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	5.0
Contribution to Project:	Overseeing all aspects of the project with the PI and leading the computational analysis of the data
Funding Support:	DOD IDA, NIH/NHLBI, Industry funds

Name:	*Martine Dumas*
Project Role:	Study Coordinator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	2.0
Contribution to Project:	Consenting patients and sample collection and supervision of IRB team.
Funding Support:	DOD IDA, NIH/NCI, Industry funds

Name:	*Robert Terrano*
Project Role:	Study Coordinator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	1.0
Contribution to Project:	Consenting patients and sample collection
Funding Support:	DOD IDA, NIH/NCI, Industry funds

Name:	*Grant Duclos*
Project Role:	Graduate Student
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	6.0
Contribution to Project:	Collecting samples, sorting cells, and preparing RNA sequencing libraries from the cells, and data analysis
Funding Support:	DOD IDA, NIH/NHLBI

Name:	*Yaron Gesthalter*
Project Role:	
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	
Contribution to Project:	Consenting patients and collection of samples
Funding Support:	Internal funds

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

J. Beane, PI
New Research Support
NIH/NCI U01CA196408 (A. Spira, PI)
Respivert Ltd. Industry award (A. Spira, PI)

Completed Research Support
American Thoracic Society/LUNGevity Foundation 2012-01 (J. Beane, PI)

What other organizations were involved as partners?
Organization Name: Broad Institute
Location of Organization: Boston, MA
Partner's contribution to the project: Collaboration. We have been collaborating with Itai Yanai, Associate Professor at Technion – Israel Institute of Technology during his sabbatical at the Broad Institute to help develop our single cell

sequencing library preparation protocol.

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Nothing to Report

What other organizations were involved as partners?
Organization Name: Broad Institute
Location of Organization: Boston, MA
Partner's contribution to the project: Collaboration. We have been collaborating with Itai Yanai, Associate Professor at Technion – Israel Institute of Technology during his sabbatical at the Broad Institute to help develop our single cell sequencing library preparation protocol.

SPECIAL REPORTING REQUIREMENTS
None.

APPENDICES:

References

1. Franklin, W. A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J. Clin. Invest.* **100,** 2133–2137 (1997).
2. Wistuba, I. I. *et al.* Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.* **89,** 1366–1373 (1997).
3. Tang, X. *et al.* EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res.* **65,** 7568–7572 (2005).
4. Mao, L. *et al.* Clonal genetic alterations in the lungs of current and former smokers. *J. Natl. Cancer Inst.* **89,** 857–862 (1997).
5. Powell, C. A., Klares, S., O'Connor, G. & Brody, J. S. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. *Clin. Cancer Res.* **5,** 2025–2034 (1999).
6. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 10143–10148 (2004).
7. Beane, J. *et al.* Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* **8,** R201 (2007).
8. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* **13,** 361–366 (2007).
9. Beane, J. *et al.* A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prev Res (Phila)* **1,** 56–64 (2008).
10. Silvestri, G. A. *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N. Engl. J. Med.* (2015). doi:10.1056/NEJMoa1504601
11. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2,** 666–673 (2012).
12. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Meth* **9,** 72–74 (2012).
13. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat Meth* **2,** 731–734 (2005).
14. Hegab, A. E. *et al.* Isolation and in vitro characterization of Basal and submucosal gland duct stem/progenitor cells from human proximal airways. *Stem Cells Transl Med* **1,** 719–724 (2012).
15. Whitney, D. H. *et al.* Derivation of a bronchial genomic classifier for lung cancer in a prospective study of patients undergoing diagnostic bronchoscopy. *BMC Med Genomics* **8,** 18 (2015).
16. Bhatia-Gaur, R. *et al.* Roles for Nkx3.1 in prostate development and cancer. *Genes Dev.* **13,** 966–977 (1999).